

Masintõlge I

Heiki-Jaan Kaalep, TÜ

Ettevaatust!

- Keegi ei tea, kuidas masintõlget teha
- Millised on tupikteed?
- Ootused on kõrgel
- Suured keeled: inglise, hiina, araabia
- Hiigelinvesteeringud uuringutesse ja tehnoloogiasse USA-s

- Euroopa vastus:
 - keeleressursid, avatud tarkvara

Projekti kokkuvõte

- Projekti fookus:
 - statistiline masintõlge
 - eesti-inglise tõlkimissuund
- Täitjad:
 - Mark Fišel
 - Heiki-Jaan Kaalep
 - Kadri Muischnek
 - Kaarel Veskis

Projekti senised tulemused

- Olemasolevad ressursid
- Ressursside kombineerimine ja kvaliteedi parandamine
- Automaatne leksikoni ekstraheerimine
- Masintõlke katsed

Olemasolevad ressursid

- Õigustekstid

- TÜ paralleelkorpus: 8 mln sõna inglise, 5 mln – eesti keeles
- Ispra JRC-Acquis: 55 mln sõna inglise, 40 mln – eesti keeles
- korpused osaliselt kattuvad

- Muu

- Opus korpus: tarkvara dokumentatsioon, subtiitrid, Euroopa konstitutsioon
- u. 1 mln sõna mõlemas keeles

Ressursside kombineerimine

- Sõltuvalt kasutatud meetodist eri korpustes 85-95% korrektsest paralleeliseid lauseid
- Välja töötatud automaatse kombineerimise meetod
 - arvestab väikeste erinevustega korpuste lausetes
 - arvestab korpuste erinevate paralleeliseamise tasemetega
- Tulemuseks saadud korpus korrektsem ja suurem kui algkorpused

Kombineeritud TÜ ja Ispra korpused

- Kombineerimise meetod rakendatud TÜ ja Ispra õigustekstide korpusetele
 - Ispra korpuse vanem versioon (7 mln sõna inglise, 5 mln – eesti keeles) – korrektsem paralleelistus
 - kattuv osa on 38% TÜ ja 25% Ispra korpusest
 - kombineeritud korpus on 193% TÜ ja 161% Ispra korpusest

Automaatne leksikon

- Inimese kasutamiseks mõeldud leksikone ei saa kasutada masintõlkes
 - tavaliselt ei ole koos mõne väljendiga ära toodud selle muutmisviisid; seetõttu ei õnnestu selle automaatne äratundmine ja kasutamine

lahendus: genereeri muutevormid

- tavapärased ja pool-produktiivsed väljendid on puudu

lahendus: tuleta leksikon automaatselt paralleelkorpuse põhjal

Automaatne leksikon

- On olemas keelest sõltumatud leksikoni tuletamise statistilised meetodid
 - rakendatud eraldi TÜ ja Ispra korpustele, et luua seadustekstide keelt kajastavat leksikoni
- Loodud ESTERM terminibaasi versioon, mille väljendite kuju teeb nende otsimise ja õige vaste leidmise lihtsamaks (1 miljon kirjet)
- Aga kas neist masintõlkes kasu on?

Masintõlke katsed

- Katse tõlkida EL seadusandlikke tekste eesti keelest inglise keelde
 - kasutamata mingit keelespetsiifilist infot peale paralleelkorpuse (eraldi TÜ ja Ispra)
 - statistilise masintõlke süsteem Moses (avatud lähtekoodiga tarkvara)
- 30% testlausetest said aktsepteeritavalt tõlgitud

Väljundi näide

Originaal:

[euroopa majandusühenduse ja Šveitsi konföderatsiooni]₁
[vaheliste kokkulepete]₂ [kohaldamisel]₃
[rakendatakse ühenduses]₄ [ühiskomitee]₅ [otsust nr 5 / 81]₆

Süsteemi tõlge:

[the european economic
community and the swiss
confederation]₁
[of agreements between]₂
[the application]₃
[of the joint committee]₅
[shall apply in the community]₄
[decision no 5 / 81]₆

Inimese tõlge:

[for the purposes of application]₃
[of the agreements between]₂
[the european economic
community and the swiss
confederation]₁ ,
[decision no 5 / 81]₆
[of the joint committee]₅
[shall apply in the community]₄

Edaspidine töö

- Kombineerida TÜ korpust uuema Ispra korpusega, katsetada leksikoni tuletamist ja masintõlget ühendkorpuse peal
- Masintõlge OPUS korpusega („elavam“ keel)
- Morfoloogiline ja süntaktiline analüüs tõlke kvaliteedi parandamiseks
- Eri faktoreid arvestavad fraasipõhised masintõlke mudelid (Moses)

Aitäh!